NPS55Gh73021A

NAVAL POSTGRADUATE SCHOOL

Monterey, California

DEVELOPMENT OF A MAN-TO-MAN RATING SCALE FOR EVALUATING PERFORMANCE

by

William H. Githens

Richard S. Elster

February 1973

NAVAL POSTGRADUATE SCHOOL
Monterey, California

Rear Admiral M. B. Freeman
Superintendent

M. U. Clauser
Provost

ABSTRACT

Over the years a continuous problem with performance rating systems has been the leniency and the non-comparability of marks assigned by different evaluators. By utilizing a computer, a method has been developed which overcomes this problem. In this new method the evaluators must compare their ratees to other specific ratees (anchors) who are under other evaluators. All ratees receive their scale value based on their relative position to the anchoring points that were used by the evaluator.

A trial of the method was made on ten groups, each composed of 12 graduate students. Each group has two evaluators. The characteristics rated were Industry, Academic Ability, Judgment, and Cooperation. These were considered as relevant characteristics to the "job" of being a student. The groups had been relatively intact for approximately one year prior to the evaluations.

A comparison of the results of using the standard fitness report rating method (RAW) and the man-to-man method (MM & MMQ) revealed that the man-to-man method was superior based on certain statistical qualities. The man-to-man method resulted in a greater spread of scores and, more importantly, resulted in higher inter-rater agreement than the standard rating method.

An outside criterion of Quality Point Average was available for the "Academic Ability" scale. Both the standard methodology and the man-to-man methodology produced rating values which were highly related to this outside criterion, .68 and .71 respectively.

This task was supported by: Chief of Naval Personnel, Personnel Research

Abstract

Development of a Man-to-Man Rating Scale for Evaluating Performance

William H. Githens
Richard S. Elster
U. S. Naval Postgraduate School

Over the years a continuous problem with performance rating systems has been the leniency and the non-comparability of marks assigned by different evaluators. By utilizing a computer, a method has been developed which overcomes this problem. In this new method the evaluators must compare their ratees to other specific ratees (anchors) who are under other evaluators. All ratees receive their scale value based on their relative position to the anchoring points that were used by the evaluator.

A trial of the method was made on ten groups, each composed of 12 graduate students. Each group has two evaluators. The characteristics rated were Industry, Academic Ability, Judgment, and Cooperation. These were considered as relevant characteristics to the "job" of being a student. The groups had been relatively intact for approximately one year prior to the evaluations.

A comparison of the results of using the standard fitness report rating method (RAW) and the man-to-man method (MM & MMQ) revealed that the man-to-man method was superior based on certain statistical qualities. The man-to-man method resulted in a greater spread of scores and, more importantly, resulted in higher inter-rater agreement than the standard rating method.

An outside criterion of Quality Point Average was available for the "Academic Ability" scale. Both the standard methodology and the man-to-man methodology produced rating values which were highly related to this outside criterion, .68 and .71 respectively.

I. INTRODUCTION

This research represents an attempt at developing and evaluating a new method for assessing the performance of U. S. Navy officers. There is a good reason for the method not having been previously developed - it would be impracticable were it not for the availability of modern digital computers. The method presented here is presumably applicable to many jobs in the civilian sector, although the work to be presented has all been conducted with Naval officers.

The U. S. Navy evaluates the performances of its officers by means of a rating form called the Report on the Fitness of officers. Over the years a continuous problem concerning Navy officer fitness reports has been the skewness of marks. The majority of officers are on the upper end of any fitness scale. Table I contains some data from an 8% sampling of U. S. Navy officer fitness reports completed in 1965. Attempts to obtain a wider or more normal distribution of marks on these scales have been given considerable attention, but have not resulted in significant improvement. These data illustrate the skewness of the distribution of marks.

TABLE 1

Ratings on "Performance of Assigned Duties". Data are from an 8% Sampling of Fitness Reports Written in 1965.[a]

Officer Grade	Outstanding High	Outstanding Low	Excellent High	Excellent Low	Very Good High	Very Good Low	Satisfactory	Inadequate
Capt.	149	91	20	4	2	0	0	0
Cdr.	287	203	61	12	3	3	4	0
LCdr.	370	324	139	57	19	8	7	1
Lt.	407	492	308	103	32	16	15	2
Lt (jg)	269	578	522	277	96	47	36	4
Ens.	63	231	364	289	135	35	32	3
Total	1545	1919	1414	742	287	109	94	10

N=6, 120

a. Source - Unpublished internal NAVBUPERS study dated 1965.

One solution frequently proposed is that the raters be forced to distribute their marks over the entire scale. Although this is desirable for certain purposes it has not been considered appropriate because even if detailing on the basis of ability were random, commands with all high or all low ability officers would occur by chance. The situation is made more severe because of selective detailing which in some cases purposefully distributes officers with higher abilities to certain assignments in rather small commands. In these cases a forced distribution would be introducing inequities into the system of evaluation by requiring that some high quality officers be given lower marks merely because of the select group to which they happen to be assigned. The man-to-man rating scheme described here should overcome some of the skewing problem without requiring a forced distribution.

II. METHOD

The man-to-man rating method proceeds as follows: Each reporting officer ranks the officers that report to him within a list of officers ("comparison" officers) of the same rank that he has known within the past three years. By "officers he has known" is meant officers he has been in charge of or officers whose work performance he has observed but who are not necessarily under his present jurisdiction. The resultant information from these ratings is then processed by a computer which considers information submitted by all raters. In this way ratings of individuals rated by more than one rater ("comparison" officers) can be averaged and used to define the value of their location (anchoring value) on the scale. A scale value is then calculated for each officer (ratee) by comparing his location on the scale with the anchoring values of the "comparison" officers. For example, if an officer is rated midway between two anchoring points, one having a computed anchor value of 4.3 and the other having a computed anchor value of 5.3, the value assigned to the officer (ratee) would be the average or 4.8.

Using the method described, experimental rating data were gathered on a population of officer students at the U. S. Naval Postgraduate School.

The following instructions were developed for the purpose of data collection.

1. We are conducting an experimental pilot study of a new method of obtaining fitness marks that was briefly described in class MN 3110. As an "experimental" study, specific names will be required, but we guarantee that information collected will be used for research purposes only.

2. We request that you assume all the men on the attached list are under your command and you must submit evaluations on their performance. You are to rate your men using the method to be described.

3. The rating method to be used does not differ conceptually from the current instructions for the fitness report. Current instructions are:

> "All evaluations made in this report shall be in comparison with officers of the same grade . . . whom you have known."

This "comparison with others" basis for the ratings means others currently in the same category (grade); it does not mean in comparison to others who were in the category at some previous time. For the purpose of this study, the subject population consists of only those men in management sections that will graduate this month. An attached list contains a group of approximately ten men to be considered as "your men." The remaining students in your section and all other MN sections comprise the "comparison group." You are to rate the performance of your men (as students) during the past year.

4. Make your ratings using the above rating concepts. However, you are to be much more explicit as to what "others" you have in mind when you rate "in comparison to others." You are to name these "comparison" officers and place them on the same scale you use to rate your men. Assume you are going to rate your men on a 15-point scale called "Loyalty."

Step 1. Think of some officer in the comparison group who is more loyal than any of your men and place the last name and initials of this "comparison officer" at the scale position best reflecting his loyalty. (In the example which follows, this officer is named "Alpha").

Step 2. Think of some current officer who is less loyal than any of your men and place the last name of this "comparison officer" on the scale. (In the example which follows, this officer is named "Beta").

Step 3. Now think of at least two more "comparison officers" and place them at the points you consider to be appropriate for them (Gamma and Delta in the following example). Circle the names of these comparison officers so they will not be confused with your men.

Step 4. Now consider the loyalty of your men one at a time and place them on the scale in relation to the men already on the scale. Adjust ratings as necessary so that your men are correctly placed relative to each other. Ties are permitted, but none of your men may tie or exceed the poorest and best officer of the "comparison group."

The scale may now look like - -

Satisfactory	1	
	2	Beta
	3	Zeta
	4	
Very Good	5	Mu
	6	
	7	Delta, Lambda
	8	
Excellent	9	Eta, Kappa
	10	
	11	Epsilon, Theta
	12	
Outstanding	13	Gamma
	14	Iota
	15	Alpha

5. You are to rate your men on four separate scales which follow. For each scale choose your "comparison officers" and rate all your men before going on to the next scale. Be sure to use only the method herein described.

The qualities chosen for rating were:

INDUSTRY: The zeal exhibited and energy applied in the per formance of his duties.

ACADEMIC ABILITY: His ability to do well scholastically in a classroom situation.

JUDGMENT: His ability to develop correct and logical conclusions.

COOPERATION: His ability and willingness to work in harmony with others.

Except for the ACADEMIC ABILITY, all the above qualities are included in the present Report of the Fitness of Officers (NavPers form 1611/1). It was felt that all four of the above qualities were relevant for performance in an academic situation and were qualities that the raters would feel they were able to use in rating fellow students.

III. POPULATION

The population studied consisted of the student/officers who graduated from the management curriculum at the Naval Postgraduate School in December of 1970. The ratings were gathered from these students in December, at the

completion of their one year assignment to the Postgraduate School. For the purpose of this study, the five sections were each randomly divided into two sub-sections. This resulted in ten sub-sections, each with approximately twelve students. Each sub-section was treated as a separate command. Each command had its own rater, which conforms to the current operational fitness report situation. A complete set of ratings consisted of one rater for each section (for a total of 10 raters), with every officer being rated by a rater. A second complete set of ratings was gathered by having a second rater for each sub-section independently develop another set of ratings. Two sets of ratings were gathered in order to study inter-rater agreement.

IV. RESULTS

There are two statistical characteristics which are necessary (necessary as distinguished from sufficient) for a good rating program. One of these is that the program produce a distribution of ratings such that there is ample differentiation between ratees on the rating scale. The other important statistical characteristic of a good rating program is that the degree of inter-rater agreement should be high. The data gathered in the first trial of the method permit an investigation of the issue of inter-rater agreement and of the characteristics of the distributions of the ratings.

A. Anchor Points

The man-to-man rating procedure depends on the characteristics of the anchor points. The anchor points are a conceptually unique feature of the method. Conceptually, the more stable (across raters) are the anchoring points, the better will be the resultant scaling. The computer program developed to implement this scaling procedure prints out a list of the anchoring points (ratees) along with their anchoring value (average ratings) and their standard deviations (across raters). The standard deviation is of special interest, for increases in its magnitude are associated with increasing differences among raters' ratings of the anchoring point (the comparison officer). Conversely if the standard deviation has a low value it indicates considerable interrater agreement. An anchoring point with a small standard deviation is better for scaling purposes than one with a large standard deviation. In an operational system a potential anchoring point's standard deviation would have to be less than some preestablished value before it would be used to influence marks assigned to any ratee. Table II contains a listing of the comparison officers, and their anchoring values along with their associated standard deviations.

TABLE 2

Frequency Distribution of the Standard Deviation of the Anchor Points (Comparison Ratees)

Range of Standard Deviations	Rater Set A Scale Number[a]				Rater Set B Scale Number[a]			
	#1	#2	#3	#4	#1	#2	#3	#4
0.0 - 0.4	1	6	3	1	2	4	3	6
0.5 - 0.9	6	10	7	7	6	17	6	3
1.0 - 1.4	6	7	4	7	6	12	9	6
1.5 - 1.9	7	3	5	6	8	1	6	3
2.0 - 2.4	2	2	4	4	2	1	6	7
2.5 - 2.9	1	3	2	1	4	2	3	4
3.0 - 3.4	4	0	2	4	4	0	0	5
3.5 - 3.9	2	1	2	1	2	2	2	1
4.0 - 4.4	1	1	0	2	1	0	2	0
4.5 - 4.9	0	0	1	1	1	0	1	3
5.0 - 5.4	0	0	0	1	1	0	0	0
5.5 - 5.9	0	0	0	0	0	0	0	1
6.0 - 6.4	1	0	0	0	0	0	0	0
Totals	31	33	30	35	37	39	38	39

a. Scale 1 - Industry, Scale 2 - Academic ability, Scale 3 - Judgment, Scale 4 - Cooperation

With the data available, it was possible to conduct scalings using: (1) the raw ratings (regular method-as if anchor points were not gathered)(2) the ratings generated using the man-to-man method; and (3) ratings generated using the man-to-man method when poor quality (high variance) anchoring points were eliminated (Qualified man-to-man method). The difference between the scaling obtained when all anchors were used and the scaling obtained after eliminating the anchors having the higher variabilities can be used to determine how sensitive the scaling procedure is to "anchor quality." In an operational-on-going application of the man-to-man methodology, the point at which an increase in the standard deviation (inter-rater disagreement) increases error variance more than it contributes to valid variance would be empirically determined. In this study an arbitrary decision was made to eliminate anchoring points which had a standard deviation greater than 3.00. This point was chosen after visually inspecting the distribution of anchoring standard deviations so that a point could be chosen which would eliminate the anchoring points with the higher standard deviations but keep the bulk of the anchors. In Scale #1 of Rater Set B, five of the 37 anchor points were eliminated using the standard deviation greater than 3.00 criterion.

B. Comparisons of the Results of the Three Scaling Methods

Several measures can be used to examine the effects and efficacies of the three scaling methods (regular, man-to-man, and qualified man-to-man). Among these measures are statistics describing the distributions of ratings obtained, the inter-rater agreement associated with each scaling method, the relationship of the resultant scales with outside criteria, and the inter-correlations among all the rater sets, traits, and methods (Multi-method--multitrait analysis). This section of the report compares the rating methods by means of the aforementioned measures.

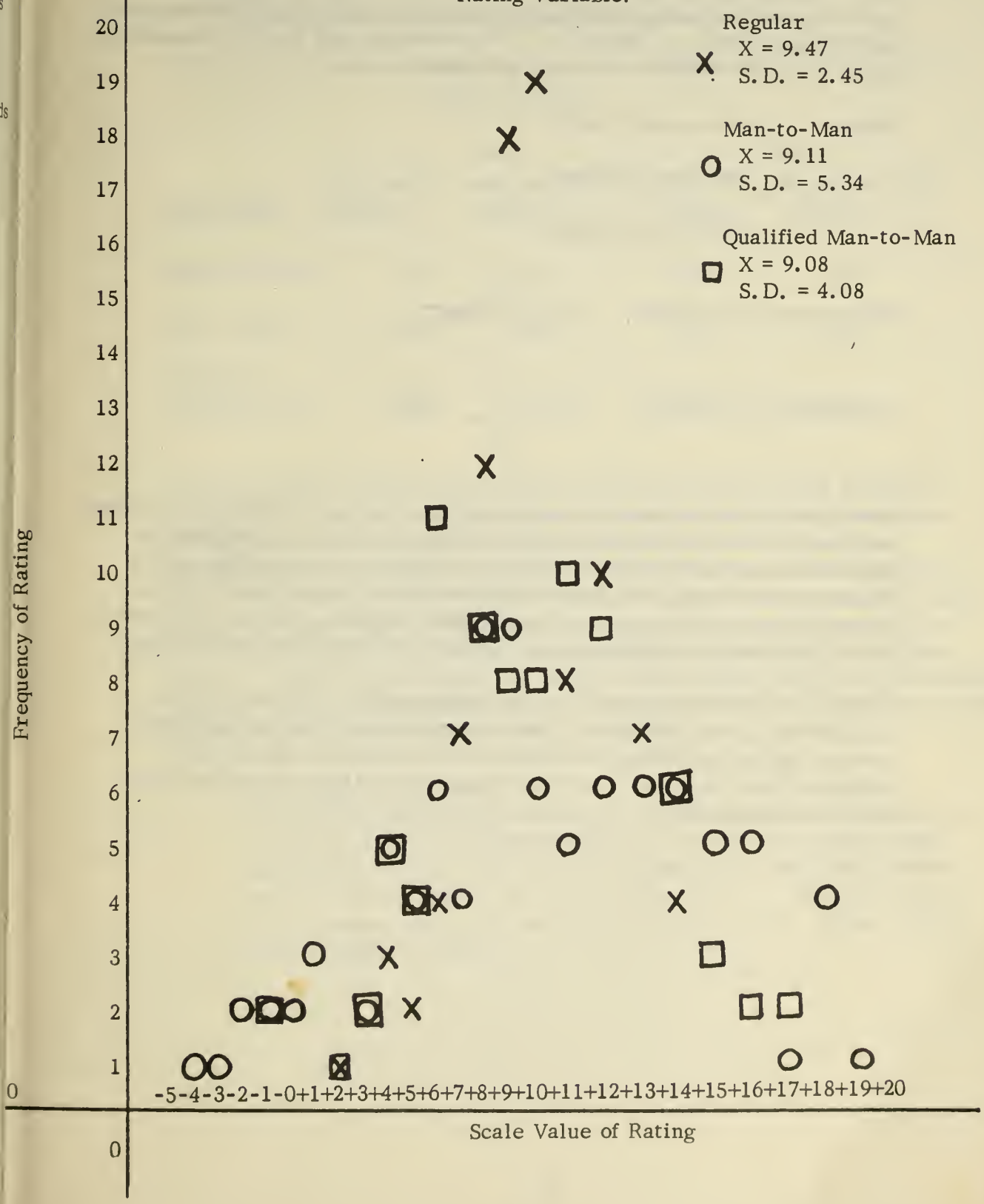
This comparison should be a severe test of the Man-to-Man methodology. In this case the Standard or Raw Method which took at face value the numerical value of the ratee's placement on the rating scale (as if comparison officers or others were not included on the same scale) has an advantage not usually associated with it. The input data (ratings) were obtained in a fashion (forcing relative comparisons between ratees as required in the Man-to-Man methodology) which should tend to increase discrimination between ratees.

1. Distribution of the Ratings

Figure 1 illustrates the distributions of ratings obtained from the three scaling methods for scale 1 and Rater Set B. The set of three distributions in Figure 1 is similar to the seven other such sets of distributions (4 scales x 2 rater sets less the set displayed in Figure 1) obtained in this study.

Figure 1

Distributions of Ratings For Scale 1, Rater Set B, Obtained When Scaling The Same Ratees Using the Following Scaling Methods: Man-to-Man, Regular, and Qualified Man-to-Man



In addition to describing the distributions shown in Figure 1, by their means and standard deviations, it is interesting to compare the forms of these distributions with that of the normal distribution. In order to compare the "shapes" of the distributions with that of the normal distribution, symmetry and kurtosis statistics were computed using the equations given in McNemar (1962, pp. 26 and 78). These indexes call for the computation of the first four moments about the distribution mean.

When - U_2 = the second moment

U_3 = the third moment

U_4 = the fourth moment

The measure of skewness $$G_1 = \frac{U_3}{U_2 \sqrt{U_2}}$$

The measure of kurtosis $$G_2 = \frac{U_4}{U_2^2} - 3$$

When both the indexes are zero, it indicates a normal distribution has been obtained. When the kurtosis statistic is less than zero, it indicates the distribution is somewhat flat-topped, and when it is greater than zero it is peaked with higher tails than those found with a normal distribution. When the index of skewness yields a positive number, the curve is skewed to the right, and a negative number indicates a skewed-left distribution.

Table 3 contains the kurtosis and symmetry statistics for each of the distributions given in Figure 1, and shows the results of the statistical tests to determine if these statistics were statistically significantly different from those that would have been obtained when sampling from normal populations.

TABLE 3

Symmetry and Kurtosis Statistics and Associated Statistical Tests for the Distributions shown in Figure 1

Rating Variable:

Distribution	Symmetry Statistic	t-test associated with symmetry	Kurtosis Statistic	t-test associated with Kurtosis
Regular Method	-.394	.112	.285	.561
Man-to-Man	-.383	.122	-.349	.476
Qualified Man-to-Man	-.360	.146	-.275	.574

An examination of the results in Table 3 which examines Scale Number 1, reveals that there is an improvement in skewness (reduction of) when the ratings used in the regular method are subjected to the Man-to-Man methodology, and even more improvement when subjected to the Qualified Man-to-Man methodology. None of the distributions are statistically significantly different from a normal distribution. For most administrative purposes it is desirable to have the distribution of ratings be flat rather than peaked. Both the Man-to-Man and Man-to-Man Qualified methodologies resulted in flatter distributions (kurtosis being -.349 and -.275) than that obtained using regular methodology (+.285).

2. Inter-Rater Agreement

The scaling data were examined in order to determine the degree to which inter-rater agreement existed. To conduct the inter-rater agreement analysis, the two sets of ratings that had been obtained were designated rater set "A" and rater set "B." The two sets of ratings were intercorrelated for each rating scale. These results are contained in Table 4.

TABLE 4

Inter-Rater Agreement:
Correlations of Rater Set A with Rater Set B[a, b]

Scale	Regular Method	Man-to-Man Method	Qualified Man-to-Man Method
1	.33	.66	.59
2	.68	.60	.71
3	.42	.37	.48
4	.21	.32	.18

a. Rater sets were formed by having a group of raters (Set A) rate the ratees and then having a completely new set of raters (Set B) rate the same ratees.

b. Data in the table are Pearson product-moment correlation coefficients.

Both the Man-to-Man method and the Man-to-Man Qualified method resulted in higher inter-rater agreement than the regular method. In the case of the Qualified Man-to-Man method the inter-rater agreement correlations were higher on all scales except the fourth. It had been anticipated that the Qualified Man-to-Man method would produce higher inter-rater agreement than the Man-to-Man method. In general, it did not do so. This may be the result of the arbitrary picking of +3 S. D. as the criterion for eliminating unreliable anchor points. In any case, further study is needed in order to understand their influence of raising or lowering the criterion for eliminating the weaker anchor points.

3. Relationships of the Three Types of Scaling to an Outside Variable

For one of the rating scales, academic ability, an external criterion was available, because a quality point ratio (QPR) was available for each of the ratees. QPR reflects academic success based upon course grades during the subject's first year of graduate work. It in turn is influenced, presumably, by academic ability--and other factors. The relationships

between QPR and the data obtained from each of the three scaling methods thus provide some indication of the validity associated with each of the scaling methods. The figures shown in Table 5 are the correlations (validity coefficients) obtained from this analysis.

TABLE 5

Correlations Between Quality Point Averages and the Ratings of Academic Ability Resulting From the Three Scaling Methods

	Scaling Method		
	Regular	Man-to-Man	Qualified Man-to-Man
Rater Set A	.69	.59	.73
Rater Set B	.73	.67	.68

Using this criterion for evaluating the methods, there is no practical difference between the Qualified Man-to-Man method and the Regular method. The Man-to-Man method resulted in a poorer showing than either the Regular or Man-to-Man Qualified methods.

4. Multi-Method & Multi Trait Analysis

Campbell and Fisk (1959) have described a method for examining the validity of psychological measures. The general logic of their scheme involves statistical methods for the construct validation of a psychological concept. The steps and logic of construct validation using the methods of Campbell and Fisk (1959) are as follows:

1. Convergent validity: Correlations between the same traits as rated by different raters are significantly different from zero.

2. Discriminant validity:

   a. The correlation between the same traits as rated by different raters should be higher than the correlation between different traits rated by the same rater.
   b. The correlation between the same traits as rated by different raters should be higher than the correlation between different traits rated by different raters.
   c. It is desirable that the same pattern of trait interrelationships

should occur in the triangles where the same rater is rating the different traits and the different raters are rating the same traits. (An example of this is that the correlation between trait A and trait B for rater 1 should be the same as the correlation between trait A for rater 1 and trait B for rater 2.)

The basic logic here is that a rating performance along dimension A of behavior is a good measure of performance along dimension A if it agrees with other ratings of performance along dimension A, but it is not a good measure of performance of dimension A if it agrees more with measures of dimension B and C than of A.

(Korman, 1971, p. 298)

The data from this study were examined using this type of analysis. Table 6 contains the complete intercorrelation matrix.

TABLE 6

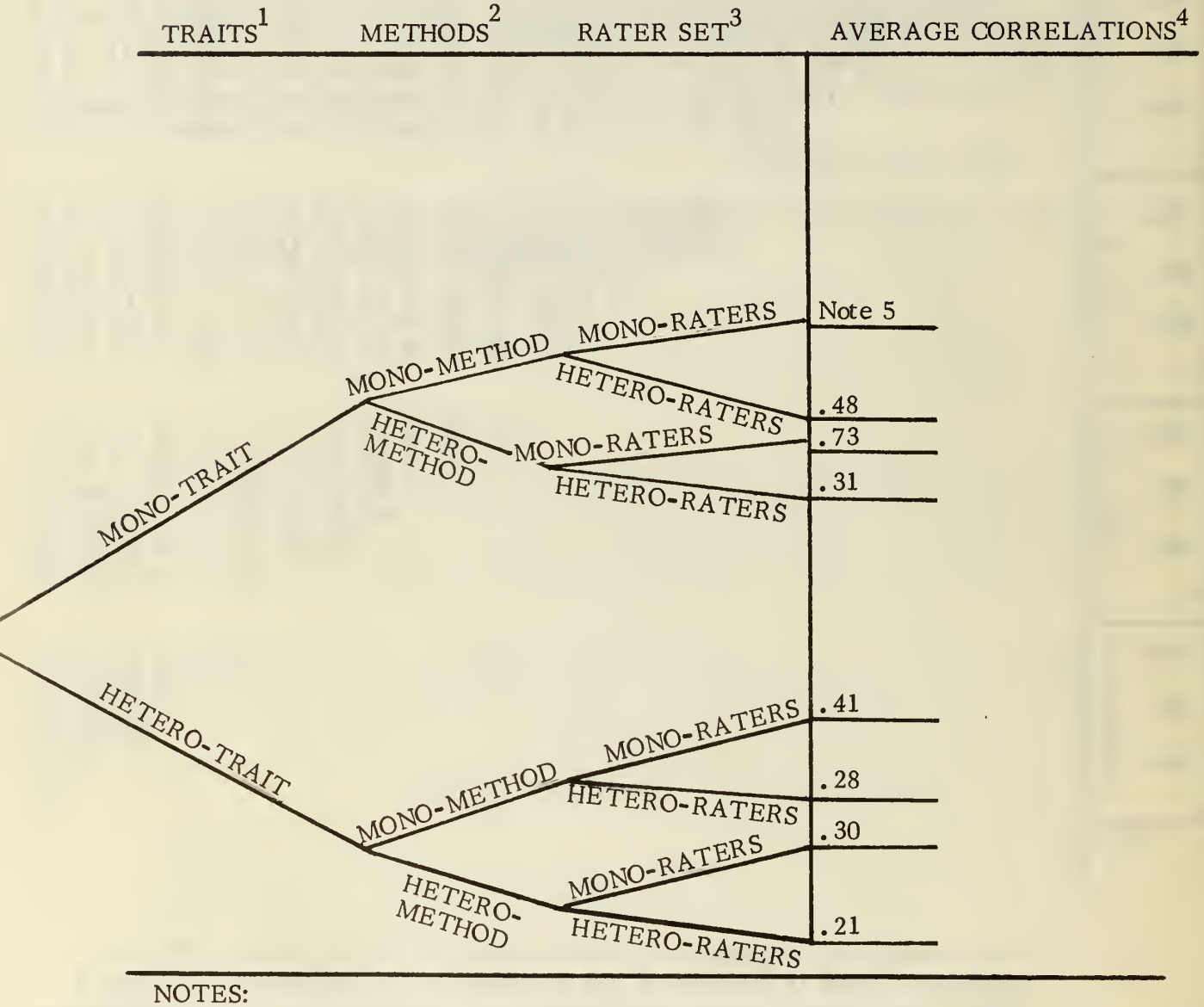
Intercorrelations Among Traits (Scales), Methods, and Rater Sets

RATER SET		SCALE #1 MMQ A	SCALE #1 MMQ B	SCALE #1 MM A	SCALE #1 MM B	SCALE #1 Regular A	SCALE #1 Regular B	SCALE #2 MMQ A	SCALE #2 MMQ B	SCALE #2 MM A	SCALE #2 MM B	SCALE #2 Regular A	SCALE #2 Regular B	SCALE #3 MMQ A	SCALE #3 MMQ B	SCALE #3 MM A	SCALE #3 MM B	SCALE #3 Regular A	SCALE #3 Regular B	SCALE #4 MMQ A	SCALE #4 MMQ B	SCALE #4 MM A	SCALE #4 MM B	SCALE #4 Regular A	SCALE #4 Regular B	QPR
SCALE #1	MMQ A		59	90	63	75	16	24	15	45	24	23	04	49	34	35	56	26	-10	39	27	39	42	23	-17	15
	MMQ B			40	63	38	49	23	40	52	45	18	19	46	52	36	64	17	01	37	21	45	56	15	06	18
	MM A				66	56	14	15	-01	27	12	11	00	25	11	14	46	14	-10	30	22	23	27	11	13	13
	MM B					25	38	07	12	37	10	00	03	17	36	37	64	-04	-11	19	27	24	23	12	11	01
	Reg. A						33	47	32	54	33	41	23	47	34	54	27	40	10	31	08	43	18	39	-04	32
	Reg. B							41	44	35	35	40	49	08	33	15	23	26	47	16	21	16	11	19	40	38
SCALE #2	MMQ A								71	88	64	96	71	46	52	46	36	69	56	21	13	22	17	19	18	73
	MMQ B									66	91	66	92	39	66	34	36	52	63	14	06	17	22	15	21	67
	MM A										60	81	58	56	65	67	62	56	36	29	14	41	26	12	03	59
	MM B											55	85	44	66	34	34	46	55	18	11	21	37	13	23	67
	Reg. A												68	49	51	35	36	74	58	16	21	12	18	25	19	62
	Reg. B													22	54	22	28	48	76	10	09	08	07	09	31	73
SCALE #3	MMQ A														48	65	34	75	10	13	19	27	37	45	-03	31
	MMQ B															54	77	38	57	18	45	17	44	15	14	39
	MM A																37	45	03	09	10	34	05	22	-10	27
	MM B																	15	30	38	45	36	52	03	-03	23
	Reg. A																		42	25	09	20	14	57	10	48
	Reg. B																			11	19	08	09	13	46	53
SCALE #4	MMQ A																				18	84	44	56	07	06
	MMQ B																					-02	72	15	54	03
	MM A																						32	50	04	06
	MM B																							22	41	06
	Reg. A																								21	13
	Reg. B																									16
QPR																										

As it stands, the complexity and extensiveness of Table 6 make it difficult to discuss. What is needed is a way of examining the correlations in Table 6 so that comparisons can be made between the magnitudes of (1) the correlations between different measures of the same trait and (2) the correlations obtained between different traits. To facilitate these comparisons, average correlations were computed using Fisher Z transformations. These data are presented in Table 7.

TABLE 7

Average Correlations Associated with Each Category of the Multi-Method Multi-Trait Analysis



NOTES:

1. Four traits are involved; Industry, Academic Ability, Judgement and Cooperation.

2. Three methods are involved; Regular, Man to Man, and Qualified Man to Man.

3. Two rater sets are involved; Rater Set A and Rater Set B.

4. Computed using Fisher's Z transformation.

5. Repeated ratings by the same set of raters using the same method on the same trait are not available.

Table 7 shows that:

1. The average correlations are higher when the same trait is being evaluated (mono-trait) than when different traits (hetero-traits) are being evaluated. (.48, .73, .31 vs, .28, .30. and , 21 respectively).

2. When evaluating the same trait, different raters using the same method produce higher intercorrelations (.48) than if different methods are used. (.31)

3. When using different methods to evaluate, the same trait, the average correlations produced by the same set of raters (.73) is higher than if different sets of raters are involved (.31).

4. When evaluating different traits, the same raters using the same methods produce higher intercorrelations (.41) than if they use different methods (.30).

5. When evaluating different traits, different raters using the same method produced higher intercorrelations (.28) than if different methods are used (.21).

6. When using different methods to evaluate different traits, the average correlation produced by the same set of raters (.30) is higher than if different sets of raters are involved (.21).

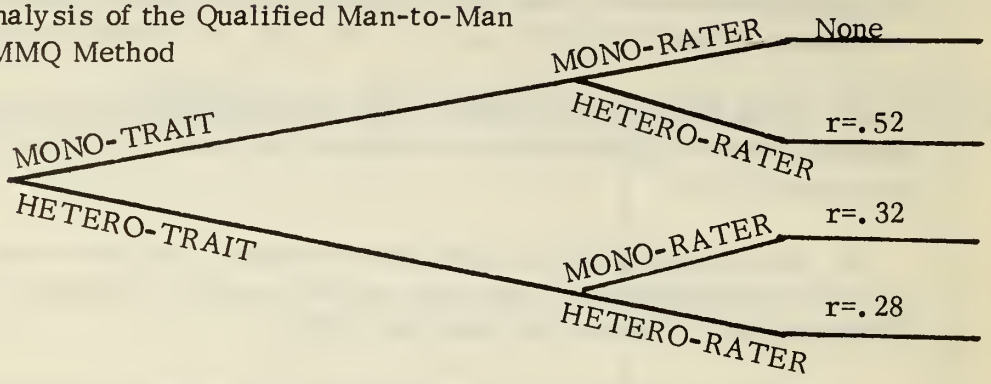
All the above relationships are in a direction that supports the validity of the traits being measured, but do not help much in a direct comparison of the three rating methodologies under study. To aid in the study of the three rating methods, a separate table (Table 8) has been generated for each rating method. The best rating methodology will be the one having the highest mono-trait correlations and the lowest hetero-trait correlations.

TABLE 8

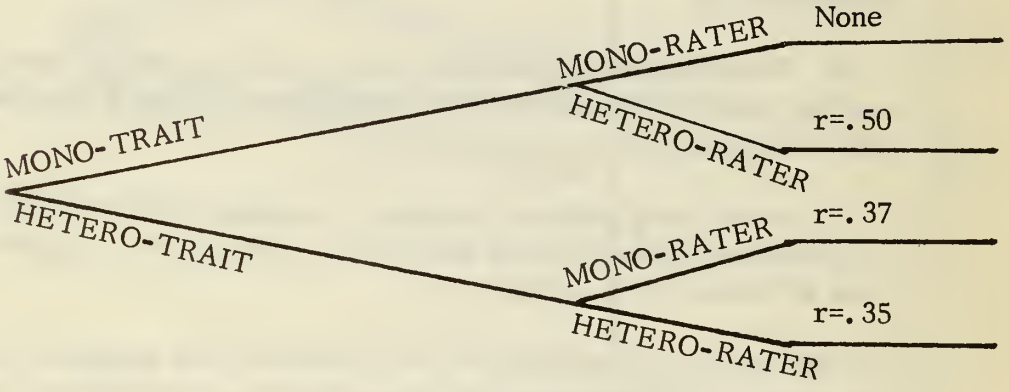
Average Correlations Associated with Each Category of a Multi-Rater, Multi-Trait Analysis

Traits[1]	Raters[2]	Average Correlations[3]

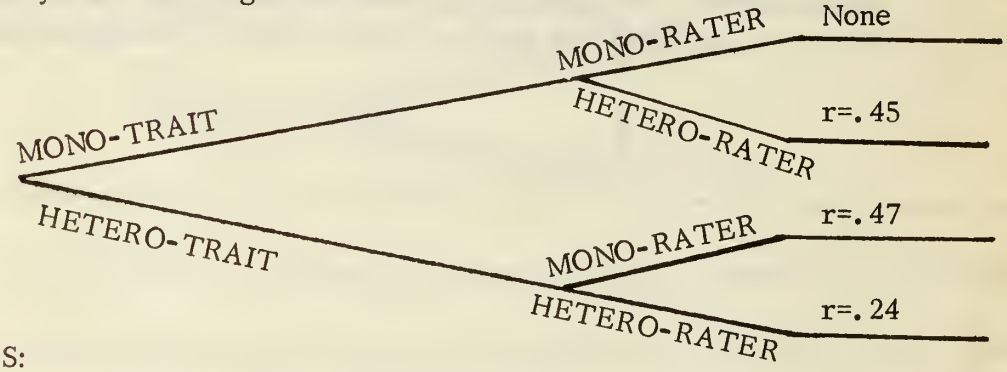
A. Analysis of the Qualified Man-to-Man (MMQ Method



B. Analysis of the Man-to-Man (MM) Method



C. Analysis of the Regular Method



NOTES:

1. Four traits are involved; Industry, Academic Ability, Judgement, and Cooperation.
2. Two rater sets are involved; Rater Set A and Rater Set B.
3. Computed using Fisher's Z transformation.

Table 8 shows that, based upon the multi-rater - multi-trait analysis, the Qualified Man-to-Man (MMQ) is the superior method for it yeilds the highest mono-trait and lowest hetero-trait correlations additionally there is a bigger difference between its mono-trait and its hetero-trait correlations than is the case with either of the other two methods. Of special note is the weakness of the Regular Method revealed by this analysis. With the Regular Rating Method, the average Mono-Trait - Hetero-Rater correlation (.45) is of approximately the same magnitude as that method's average Hetero-Trait - Mono-Rater correlation (.47).

V. CONCLUSIONS & RECOMMENDATIONS

The statistical investigations of the data obtained from the three rating methodologies have shown the man-to-man rating methodology to be somewhat superior to the regular rating methodology.

The new methodology resulted in a better (greater) distribution of ratings, more agreement among raters, and better differentiation among rating scales (Industry, Academic Ability, Judgment, and Cooperation) used in this study.

It is recommended that another set of ratings be gathered and the same evaluative procedures be used with those data. These ratings should also be gathered at two (or more) different times from the same raters in order to investigate the monorater monotrait correlations.

It is also recommended that the man-to-man methodology be used at a Naval Command to refine it further, and draw implications about its applicability to the entire Navy.

This report was begun with a general discussion of rating problems in general and stressed the tendency for raters to be lenient when rating subordinates. The method used in this study to develop anchor ratees required the rater to select anchoring ratees from outside his section. More specifically, a rater was required to name, for each rating scale, an individual from outside his section who was higher on the scale than anyone in the rater's section. Additionally, the rater was required to name another individual from outside the rater's section who was lower on the scale then anyone in the rater's section. The authors of this report recognize that, at times, a rater may consider it impossible to pick someone from outside who is better/poorer than his best/worst ratee (on some attribute), but to overcome the leniency tendency, the

rater should be urged, pushed, cajoled, etc., to try to live by the instructions for the selection of anchor men as they were used in this study. (Anecdotally, the authors noted that none of the raters had difficulty finding a comparison officer lower than any of their ratees, while some raters claimed difficulty at finding a comparison officer higher than any of their ratees. The authors conclude that the leniency effect is persistent and pervasive, when rating people in one's own group).

REFERENCES

1. Campbell, D. T. and D. Fiske, Convergent and Discriminant Validation by the Multitrait Multimethod Matrix, Psychological Bulletin, 1959.

2. Korman, A. K., Industrial Organizational Psychology, New York: Prentice-Hall, 1971.

INITIAL DISTRIBUTION LIST

Dean J. M. Wozencraft Dean of Research Naval Postgraduate School Monterey, California 93940	1
Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
Defense Documentation Center (DDC) Cameron Station Alexandria, Virginia 22314	12
Library, Code 55 Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	5
W. G. Githens, Code 55Gh Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	5
R. S. Elster, Code 55Ea Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	5

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)

Naval Postgraduate School
Monterey, California 93940

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3. REPORT TITLE

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report

5. AUTHOR(S) (First name, middle initial, last name)

William H. Githens and Richard S. Elster

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
February 1973		2

8. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY

13. ABSTRACT

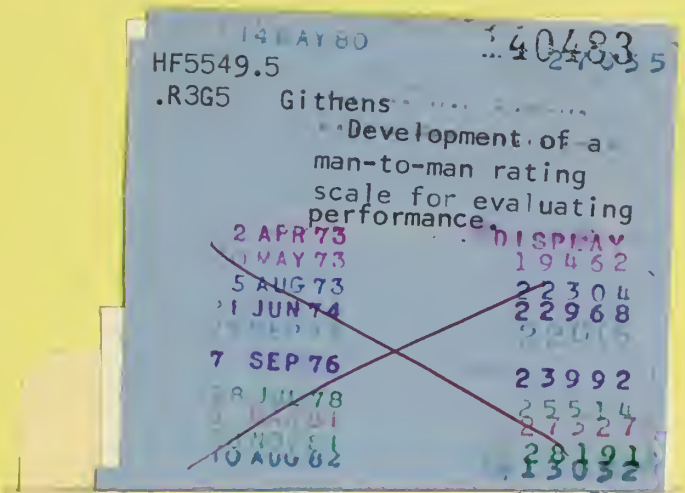
Over the years a continuous problem with performance rating systems has been the leniency and the non-comparability of marks assigned by different evaluators. By utilizing a computer, a method has been developed which overcomes this problem. In this new method the evaluators must compare their ratees to other specific ratees (anchors) who are under other evaluators. All ratees receive their scale value based on their relative position to the anchoring points that were used by the evaluator.

A trial of the method was made on ten groups, each composed of 12 graduate students. Each group has two evaluators. The characteristics rated were Industry, Academic Ability, Judgment, and Cooperation. These were considered as relevant characteristics to the "job" of being a student. The groups had been relatively intact for approximately one year prior to the evaluations.

A comparison of the results of using the standard fitness report rating method (RAW) and the man-to-man method (MM & MMQ) revealed that the man-to-man method was superior based on certain statistical qualities. The man-to-man method resulted in a greater spread of scores and, more importantly, resulted in higher inter-rater agreement than the standard rating method.

An outside criterion of Quality Point Average was available for the "Academic Ability" scale. Both the standard methodology and the man-to-man methodology produced rating values which were highly related to this outside criterion, .68 and .71 respectively.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Ratings						
Rating Scales						
Evaluation						
Fitness Reports						
Measurement						
Effectiveness						
Criteria						
Performance						
Merit Ratings						
Scaling						



140483
HF5549.5
.R3G5 Githens
Development of man-to-man rating scale of evaluating performance.